ECO 391 Economics and Business Statistics

Lecture 8: Inference with Regression Models

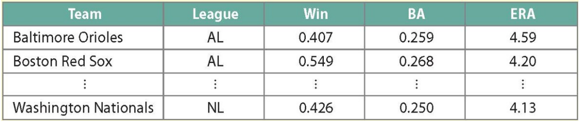
Xiaozhou Ding

March 20, 2019

Outline

- Conduct a test of individual significance.
- Conduct a test of joint significance.
- Explain the role of the assumptions on the OLS estimators.
- Describe common violations of the assumptions and offer remedies.

Introductory Case: Analyzing the Winning Percentage in Baseball



Team	League	Win	BA	ERA
Baltimore Orioles	AL	0.407	0.259	4.59
Boston Red Sox	AL	0.549	0.268	4.20
⋮	⋮	⋮	⋮	⋮
Washington Nationals	NL	0.426	0.250	4.13

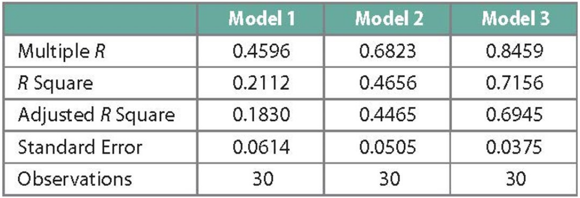
- Sports analysts frequently quarrel over what statistics separate winning teams from the losers.
- Is a high batting average (BA) the best predictor, or is it a low earned run average (ERA)? Or both?
- We will fit three regression models and use the statistical significance of the predictors to help decide.

Test of Significance

Three Models

With two explanatory variables to choose from, we can formulate three linear models:

- Model 1: Win $= \beta_0 + \beta_1 BA + \epsilon$
- Model 1: Win $= \beta_0 + \beta_1 ERA + \epsilon$
- Model 1: Win $= \beta_0 + \beta_1 BA + \beta_2 ERA + \epsilon$



	Model 1	Model 2	Model 3
Multiple *R*	0.4596	0.6823	0.8459
R Square	0.2112	0.4656	0.7156
Adjusted *R* Square	0.1830	0.4465	0.6945
Standard Error	0.0614	0.0505	0.0375
Observations	30	30	30

Economic Significance and Statistical Significance

- Economic significance: if the marginal effect of x is large, we say x is economically significant.
  - Usually we use the magnitude to determine whether the coefficient is economic significant. But notice the coefficient is not unit free, so interpreting it correctly is important.
- Statistical significance: the coefficient of x (marginal effect) is statistically different from zero.
  - Usually statistically significance can be driven by a large estimate of coefficient, or a small standard error.
- How do we determine if the coefficient of x is statistically significant?

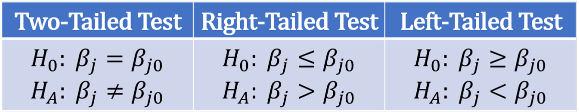
Tests of Individual Significance

We can test individual significance for both simple and multiple linear regression models.

- Consider the standard multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

- In general, we can test whether β_j is equal to, greater than, or less then some hypothesized value β_{j0}.
- This test could have one of three forms:



Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
H_0: $\beta_j = \beta_{j0}$ H_A: $\beta_j \neq \beta_{j0}$	H_0: $\beta_j \leq \beta_{j0}$ H_A: $\beta_j > \beta_{j0}$	H_0: $\beta_j \geq \beta_{j0}$ H_A: $\beta_j < \beta_{j0}$

Test Statistic for the Test of Individual Significance-T value

Given

- $df = n - k - 1$
- n is sample size
- k is the number of explanatory variables
- b_j is the estimate for β_j
- $se(b_j)$ is the standard error of the OLS estimator b_j
- β_{j0} is the hypothesized value of β_j

Then the value of the test statistic for a test of individual significance is

$$t_{df} = \frac{b_j - \beta_{j0}}{se(b_j)}$$

Note, $\beta_{j0} = 0$, the value of the test statistic is reduced to

$$t_{df} = \frac{b_j}{se(b_j)}$$

Testing $\beta_j = 0$

- By far the most common hypothesis test for an individual coefficient is to test whether its value differs from zero.
- Too see why, consider our model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

- If a coefficient is equal to zero, then it implies that the explanatory variable is not a significant predictor of the response variable.

Computer-Generated Output

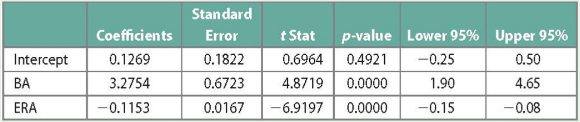
- Virtually all statistical software will automatically report a test statistic and a p-value with each coefficient estimate.
- These values can be used to test whether the regression coefficient differs from zero. $(H_0 : \beta_j = 0)$.
- To perform a one-sided test where the hypothesized value is zero, divide the computer-reported p-value in half.
- If we wish to test whether the coefficient differs from a nonzero value, we need to compute a new test statistic.

Example

- To test whether batting average influences winning percentage, we set up the following hypotheses:

$$H_0 : \beta_1 = 0; \quad H_A : \beta_1 \neq 0$$

- Then, examine the regression output.



	Coefficients	Standard Error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	0.1269	0.1822	0.6964	0.4921	−0.25	0.50
BA	3.2754	0.6723	4.8719	0.0000	1.90	4.65
ERA	−0.1153	0.0167	−6.9197	0.0000	−0.15	−0.08

- The value of the test statistic is $t_{27} = 4.817$ and its p-value is very close to zero. We reject the null hypothesis and conclude that batting average is a significant predictor, Since the critical value is $t_{\alpha/2,df} = 1.96$.

Using a Confidence Interval to Determine Individual Significance

- We can use a confidence interval to conduct a two-tailed hypothesis test to see if a regression coefficient differs from zero.
- A $100(1-\alpha)\%$ confidence interval for the regression coefficient β_j is computed as

  $$b_j \pm t_{\alpha/2,df} se(b_j)$$

  or

  $$\left[b_j - t_{\alpha/2,df} se(b_j), b_j + t_{\alpha/2,df} se(b_j)\right]$$
- Terms are the same as for the test statistic for a test of individual significance

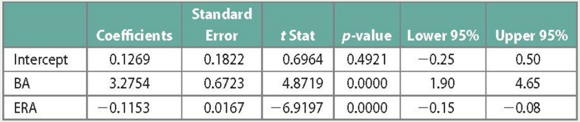
Using a Confidence Interval to Determine Individual Significance

If the confidence interval for a slope coefficient

- contains the value zero: the explanatory variable associated with the regression coefficient is not significant (fail to reject the null hypothesis)
- does not contain the value zero: the explanatory variable associated with the regression coefficient is statistically significant (reject the null hypothesis)

Example

For *ERA*, the interval of $[-0.15, -0.08]$ does not include 0, indicating *ERA* is a significant predictor.

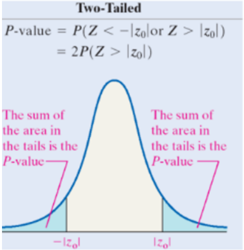


	Coefficients	Standard Error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	0.1269	0.1822	0.6964	0.4921	−0.25	0.50
BA	3.2754	0.6723	4.8719	0.0000	1.90	4.65
ERA	−0.1153	0.0167	−6.9197	0.0000	−0.15	−0.08

Using p-value

- p-value is the probability of finding the observed sample results, or ”more extreme” results, when the null hypothesis is actually true (where ”more extreme” is dependent on the way the hypothesis is tested).
- If the p-value is equal to or smaller than the significance level (α), it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true, and thus that hypothesis must be rejected and the alternative hypothesis is accepted as true.

Two-tail

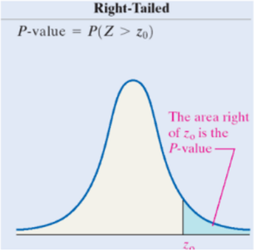


Two-Tailed

$$P\text{-value} = P(Z < -|z_0| \text{or } Z > |z_0|)$$
$$= 2P(Z > |z_0|)$$

The sum of the area in the tails is the P-value
The sum of the area in the tails is the P-value
$-|z_0|$
$|z_0|$

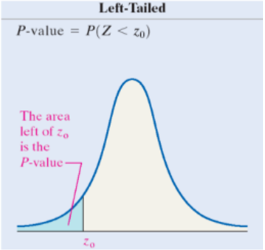
Right Tail



Right-Tailed

$$P\text{-value} = P(Z > z_0)$$

Left Tail



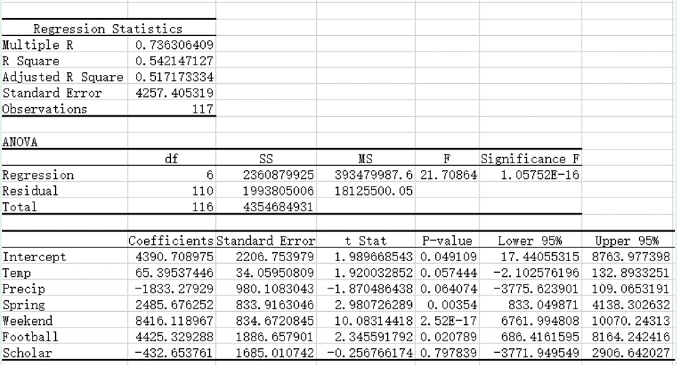
Left-Tailed

$$P\text{-value} = P(Z < z_0)$$

Example

Suppose we want to study the determinants of attendance to Keenland. We have the following regression model:

$$\begin{aligned} attendance_i = \beta_0 + \beta_1 temperature_i + \beta_2 precipitation_i + \beta_3 spring_i \\ + \beta_4 weekend_i + \beta_5 football_i + \beta_6 scholarship_i + \epsilon_i \end{aligned}$$



Regression Statistics	
Multiple R	0.736306409
R Square	0.542147127
Adjusted R Square	0.517173334
Standard Error	4257.405319
Observations	117

ANOVA

	df	SS	MS	F	Significance F
Regression	6	2360879925	393479987.6	21.70864	1.05752E-16
Residual	110	1993805006	18125500.05		
Total	116	4354684931			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	4390.708975	2206.753979	1.989668543	0.049109	17.44055315	8763.977398
Temp	65.39537446	34.05950809	1.920032852	0.057444	-2.102576196	132.8933251
Precip	-1833.27929	980.1083043	-1.870486438	0.064074	-3775.623901	109.0653191
Spring	2485.676252	833.9163046	2.980726289	0.00354	833.049871	4138.302632
Weekend	8416.118967	834.6720845	10.08314418	2.52E-17	6761.994808	10070.24313
Football	4425.329288	1886.657901	2.345591792	0.020789	686.4161595	8164.242416
Scholar	-432.653761	1685.010742	-0.256766174	0.797839	-3771.949549	2906.642027

Example

- Does “weekend” have a statistically significant effect on Keenland attendance at the 5% significance level?
  - Using the p-value
  - Using the t-statistic
  - Using the confidence interval
- Do you get consistent conclusions
- We get consistent conclusions using all three methods. The variable “Weekend” has a statistically significant effect on attendance.

Test of Joint Significance

- In addition to conducting tests of individual significance, we also may wan to test the joint significance of all k variables at once.
- The competing hypotheses for a test of joint significance are

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0, \quad H_A : \text{at least one } \beta_j \neq 0$$

- Notice: you cannot test each explanatory variable separately. You must do a joint test. **Testing a series of individual hypothesis is not equivalent to testing the same hypothesis jointly.**

The Test Statistic

- The test statistic for a test of joint signifiance is

$$F_{df_1,df_2} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$

  where MSR and MSE are the mean regression sum of squares and the mean error sum of squares, respectively.
- The numerator degrees of freedom, df_1, equal k, while the denominator degrees of freedom, df_2, are $n-k-1$.
- Fortunately, statistical software will generally report the value of F_{df_1,df_2} and its p-value as standard output, making computation by hand rarely necessary.

F distribution is another common distribution used in statistics. It is a right skewed disribution. The critical values can be found online.
`http://www.socr.ucla.edu/applets.dir/f_table.html`

Example

- We want to conduct a joint test of significance for the model

$$Win = \beta_0 + \beta_1 BA + \beta_2 ERA + \epsilon$$

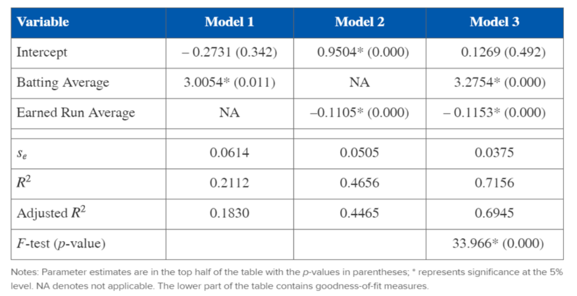
- So we set up the following hypotheses:

$$\begin{aligned} H_0 :& \beta_1 = \beta_2 = 0 \\ H_A :& \text{at least one } \beta_j \neq 0 \end{aligned}$$

- From the ANOVA portion of the regression results, we see that $F_{2,27} = 33.9963$ and its p-value is quite small, so we reject the null hypothesis, and conclude that the explanatory variables (regression) are jointly significant. ($F = 3.3541$ is the critical value under $\alpha = 0.05$ significance level.)

Example

- Goodness-of-fit measures indicated that including both batting average and ERA is most appropriate.



Variable	Model 1	Model 2	Model 3
Intercept	– 0.2731 (0.342)	0.9504* (0.000)	0.1269 (0.492)
Batting Average	3.0054* (0.011)	NA	3.2754* (0.000)
Earned Run Average	NA	–0.1105* (0.000)	– 0.1153* (0.000)
s_e	0.0614	0.0505	0.0375
R^2	0.2112	0.4656	0.7156
Adjusted R^2	0.1830	0.4465	0.6945
F-test (p-value)			33.966* (0.000)

Notes: Parameter estimates are in the top half of the table with the *p*-values in parentheses; * represents significance at the 5% level. NA denotes not applicable. The lower part of the table contains goodness-of-fit measures.

- In Model 3, explanatory variables are individually significant and the regression is jointly significant.
- We can conclude that both batting average and earned run average are good predictors of overall winning percentage.

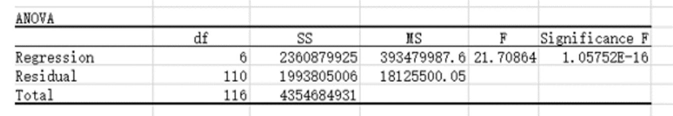
Example

Test if the independent variables in the regression of Keenland attendance are jointly significant.

1. Hypothesese

   H_0 : all slope coefficients are zero; $\quad H_A$: at least one slope coefficient is not zero

2. Test statistic:



ANOVA

	df	SS	MS	F	Significance F
Regression	6	2360879925	393479987.6	21.70864	1.05752E-16
Residual	110	1993805006	18125500.05		
Total	116	4354684931			

   - $df_1 = k = 6$, $df_2 = n-k-1 = 110$, $SSR = 2360879925$, $SSE = 1993805006$.
   - $F_{6,110} = \frac{SSR/k}{SSE/(n-k-1)} = 21.70864$.

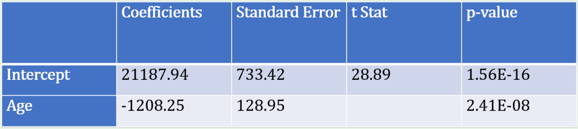
3. At the 5% significance level, the F critical value is 2.182.

4. Since our test statistic is in the reject region, we reject our null hypothesis and conclude that at least one coefficient is significant.

Note, the ANOVA table will give you the F statistic directly, as well as the corresponding p-value. From the p-value, you can conclude whether the explanatory variables are jointly significant.

Example

In order to examine the relationship between the selling price of a used car and its age, an analyst uses data from 20 recent transactions and estimates



	Coefficients	Standard Error	t Stat	p-value
Intercept	21187.94	733.42	28.89	1.56E-16
Age	-1208.25	128.95		2.41E-08

- Specify the competing hypotheses in order to determine whether the selling price of a used car and its age are linearly related.
- Calculate the value of the test statistic.
- At the 5% significance level, what is the conclusion to the test? Is the age of a used car significant in explaining its selling price?
- Conduct a hypothesis test at the 5% significance level in order to determine ifβ_1 differs from -1000. Show all the relevant steps.

Example

- Specify the competing hypotheses in order to determine whether the selling price of a used car and its age are linearly related.

$$H_0 : \beta_1 = 0; \quad H_A : \beta_1 \neq 0$$

- Calculate the value of the test statistic.
  Given $df = n - k - 1 = 20 - 1 - 1 = 18$, we calculate the value of the test statistic as

$$t_{18} = \frac{b_1 - \beta_{1,0}}{se(b_1)} = \frac{-1208.25 - 0}{128.95} = -9.37$$

- At the 5% significance level, what is the conclusion to the test? Is the age of a used car significant in explaining its selling price?
  With $\alpha = 0.05$, $t_{\alpha/2,df} = t_{0.025,18} = 2.10$. For a two-tailed test, the critical values are 2.10 and -2.10, we reject H_0. At the 5% significance level, we can conclude that $\beta_1 \neq 0$. Thus, the age of a used car is significant in explaining its selling price.

Example

- Conduct a hypothesis test at the 5% significance level in order to determine if β_1 differs from -1000. Show all the relevant steps.
  First specify the competing hypotheses:

$$H_0 : \beta_1 = -1000; \quad H_A : \beta_1 \neq -1000$$

Given $df = n - k - 1 = 20 - 1 - 1 = 18$, we calculate the value of the test statistic as

$$t_{18} = \frac{b_1 - \beta_{1,0}}{se(b_1)} = \frac{-1208.25 - (-1000)}{128.95} = -1.61$$

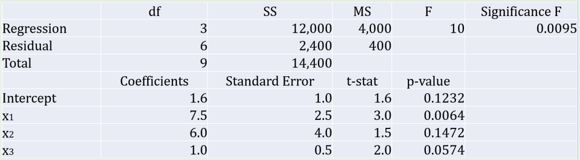
With $\alpha = 0.05$, $t_{\alpha/2,df} = t_{0.025,18} = 2.10$. For a two-tailed test the critical values are 2.10 and-2.10. The decision rule is to reject H_0 if $t_{18} > 2.10$ or $t_{18} < -2.10$. Since $-2.10 < -1.61 < 2.10$, we do not reject H_0. At the 5% significance level, we cannot conclude that $\beta_1 \neq -1000$.

Example

An analyst examines the effect that various variables have on crop yield. He estimates

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where y is the average yield in bushels per acre, x_1 is the amount of summer rainfall, x_2 is the average daily use in machine hours of tractors on the farm, and x_3 is the amount of fertilizer used per acre. The results of the regression are as follows:



	df	SS	MS	F	Significance F
Regression	3	12,000	4,000	10	0.0095
Residual	6	2,400	400		
Total	9	14,400			
	Coefficients	Standard Error	t-stat	p-value	
Intercept	1.6	1.0	1.6	0.1232	
x1	7.5	2.5	3.0	0.0064	
x2	6.0	4.0	1.5	0.1472	
x3	1.0	0.5	2.0	0.0574	

- At the 10% significance level, are the explanatory variables jointly significant in explaining crop yield? Explain.
- At the 10% significance level, can you conclude that the slope coefficient attached to rainfall differs from 9? Explain. ($t_{0.05,6} = 1.943$)

Example

- The competing hypotheses for a joint significance test of the three explanatory variables take the form: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$; H_A : at least one $\beta_j \neq 0$. Because the p-value associated with $F_{3,6} = 10$ is 0.0095 and this value is less than 0.10, we reject H_0 and conclude that at least one of the explanatory variables is significant in explaining crop yield.
- $H_0 : \beta_1 = 9$; $H_A : \beta_1 \neq 9$. The value of the test statistic is $t_6 = -0.6$. The critical value at the 10% significance level are $-t_{0.05,6} = -1.943$ and $t_{0.05,6} = 1.943$. Because $-1.93 < -0.6 < 1.943$, we do not reject H_0 and cannot conclude that the slope coefficient attached to rainfall differs from 9.

Model Assumptions and Common Violations

OLS Assumptions

The Statistical properties of the OLS estimator, as well as the validity of the testing procedures, depend on a number of assumptions. We discuss those assumptions now.

1. Linearity. The regression model given by
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$
is linear in the parameters, $\beta_0, \beta_1, \ldots, \beta_k$, with an additive error term ϵ.
2. Conditional on x, the error term has an expected value of zero, or $\mathrm{E}(\epsilon) = 0$. This implies
$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$
3. There is no exact linear relationship among the explanatory variables; there is no perfect collinearity.
4. No endogeneity. The error term ϵ is not correlated with any of the explanatory variables.(Assumption violated if important explanatory variables are excluded.)
5. Homoskedastic error. Conditional on $x(x_1, x_2, x_3, \ldots, x_k)$, the variance of the error term is the same for all observations. (Assumption violated if observations have a changing variability)
6. No serial correlation: conditional on x, the error term ϵ is uncorrelated across observations. (Assumption is violated if observations are correlated.)
7. The error term ϵ is normally distributed. This assumption allows us to construct confidence interval and conduct test of significance. If ϵ is not normally distributed, the hypothesis test is only valid for a large sample size.

OLS Assumptions

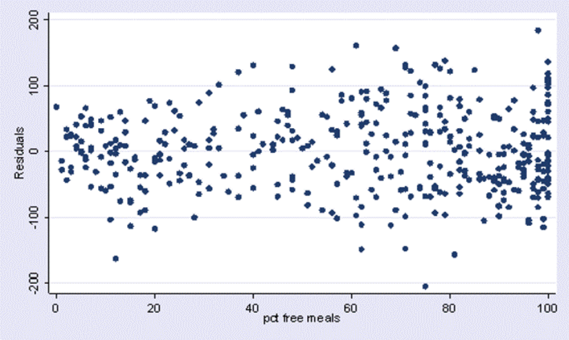
- Under the above assumptions, the OLS estimators of the regression coefficients β_j are unbiased; that is $\mathrm{E}(b_j) = \beta_j$.
- Among all linear unbiased estimators, they have minimum variations between samples.
- If one or more assumptions are violated, the OLS estimators will be compromised.
- Violation of certain assumptions also impacts the validity of the significance tests. Because the t-test and F-test rely on the standard error of the estimators.

Checking the Assumptions

- The true error terms ϵ cannot be observed because they exist only in the population. We can, however, look at the residuals, $e = y - \hat{y}$, where $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$, for each observation.
- It is common to plot the residuals on the vertical axis and an explanatory on the horizontal axis.
- When estimating a regression in Excel, the dialog box that opens after choosing Regression in Data Analysis toolbox.

Residual Plot

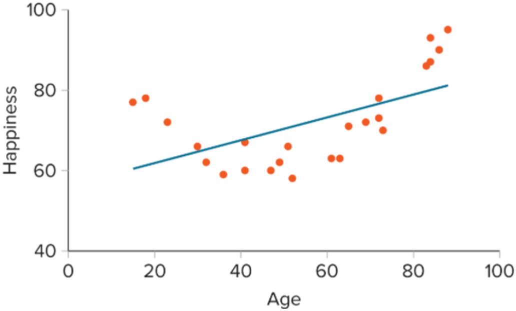
If the function is correctly specified, the residual plot should look like randomly distributed.



Common Violation 1: Nonlinear Patterns

- Simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$ implies that if x goes up by 1 unit, we expect y to change by β_1, irrespective of the value of x.
- However, if the relationship between response variable and explanatory variable cannot be represented by a straight line.
- Now, you need use your economic theory and intuition to determine if the linearity assumption is right.

Detection: Use scatter plots or residual plots to confirm your intuition.
Remedy: Use nonlinear regression methods based on simple transformation of the response and explanatory variables.



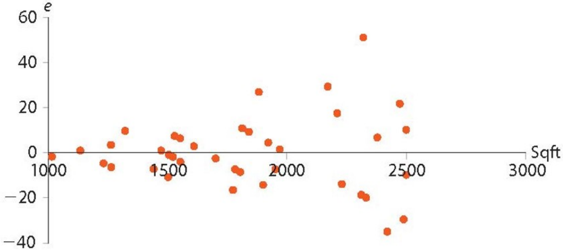
Common Violation 2: Multicollinearity

- This occurs when two or more explanatory variables have an exact linear relationship, which makes it difficult to disentangle the influence of each explanatory variable on the response, we may find some explanatory variables are insignificant or have the wrong sign.
- Detection:
  - High R^2 and individually insignificant variables.
  - Sample correlation coefficients between two explanatory variables greater than .8 or less than -.8.

  Example: $income = \beta_0 + \beta_1 male + \beta_2 female + \epsilon$
- Remedy:
  - Simply drop one.
  - Increase sample size.
  - Transform variables so no longer be collinear.
  - Last, especially if we are interested only in maintaining a high predictive power, it may make sense to do nothing.

Common Violation 3: Changing Variability (Heteroskedasticity)

- Variance of observations not constant (and thus variance of residuals not constant).
- Estimators still unbiased but standard errors incorrect, gives us meaningless F and t test results.
- Detection: Make a residual plot against all the explanatory variables or predicted values $\hat{y}$, if there is a trend we have changing variability

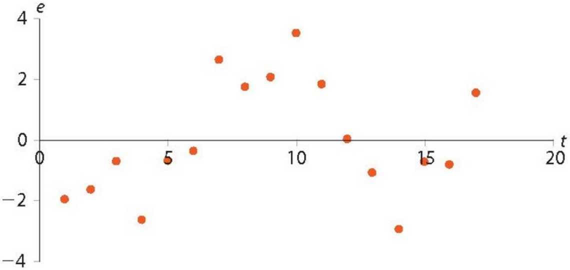


- Remedy: To get around this problem, some researchers use OLS estimates along with corrected standard errors, often referred to as robust standard errors. Many statistical packages have this option available, unfortunately the current version of Excel does not.

Common Violation 4: Correlated Observations

- We assume that the error term is uncorrelated across observations when obtaining OLS estimates.
- But this often breaks down in time series data. The error term is correlated across observations. (Often happens when you have time-series data. Variables such as GDP, employment, and asset returns are often considered serially correlated)
- Estimators still unbiased, but standard errors distorted downward leading the model to look better than it is variables may look individually and jointly significant when they are not.
- Detection: : Plot residuals against time to look for serial correlation. If the residuals show no pattern around the horizontal axis, the serial correlation is not likely to be a problem.

- For example, we predict sales at a sushi restaurant over a period of time. A plot of the residuals against time shows:



- Remedy: use OLS estimators but correct the standard errors using the Newey-West procedures.(Remedies are not easily accessible using Excel)

Common Violation 5: Excluded Variables

- Important explanatory variables are excluded, then the error term becomes correlated with both the included and excluded explanatory variables - the degree to which this occurs depends on the correlation between the included and excluded explanatory variables.
- Give biased coefficient estimates-possibly even of the wrong sign
- Remedy:
  - Good modeling techniques (considering all explanatory variables)
  - Instrumental variable approach-beyond scope of this class

Example

The return to education.

- If we run a simple linear regression model of income on education.
- This model excludes innate ability, which is an important ingredient for income.
- Since ability is omitted, it gets incorporated in the error term and the resulting error term is likely to be correlated with years of education.